DOCUMENT RESUME

ED 283 860 TM 870 384

AUTHOR	Fischl, M. A.; And Others
TITLE	Development of Officer Selection Battery Forms 3 and 4. Technical Report 603.
INSTITUTION	American Institutes for Research in the Behavioral Sciences, Washington, D.C.
SPONS AGENCY	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
REPORT NO	ARI-TR-603
PUB DATE	Mar 86
CONTRACT	MDA-903-80-C-0701
NOTE	70p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	College Students; Correlation; Ethnic Groups; Higher Education; High Schools; High School Seniors; Item Analysis; Job Analysis; *Military Personnel; Military Training; *Occupational Tests; *Personnel Selection; Predictive Validity; Sex Differences; *Test Construction; Test Items; *Test Validity
IDENTIFIERS	Cadet Corps; Military Recruitment; *Officer Selection Battery; *Reserve Officers Training Corps

ABSTRACT

This report describes the development, standardization, and validation of two parallel forms of the Officer Selection Battery, a 2-hour, group administrable, paper and pencil test for assessing men and women applying for the Reserve Officers Training Corps (ROTC). Based on an extensive job analysis, 1,400 experimental items in 12 job areas were developed and administered to 3,306 college juniors who were enrolled in the Advanced ROTC program. This group was stratified to conform to the 1980 national distribution of Scholastic Aptitude Test scores, and weighted to reflect appropriate ethnic and gender representation. Item analyses were performed using faculty evaluations of each tested cadet's officer potential as criteria. Items that would yield the most valid tests with the least gender or ethnic impact were selected for the final forms. For standardization, the tests were administered to college sophomores in military science courses. Criterion-related validity was investigated twice using samples of senior ROTC cadets and Second Lieutenants in their first assignments. On the basis of the correlation and regression analyses performed, the Officer Selection Battery was concluded to be empirically and content valid and of comparable validity for ethnic and gender subgroups, with no indication of differential validity or regressions. Appendices include definitions of identified dimensions of officer job performances, test item types, the plan for addressing equal employment fairness in the Officers Selection Battery, weighting procedures, readability measures, and a comparison of the Officer Selection Battery with two tests in use as selectors for army officer precommissioning training programs. (JAZ)

ED283860

Technical Report 603

Development of Officer Selection Battery Forms 3 and 4

M. A. Fischl
Army Research Institute

Dorothy S. Edwards and John G. Claudy
American Institutes for Research

Michael G. Rumsey
Army Research Institute

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Selection and Classification Technical Area
Manpower and Personnel Research Laboratory

U. S. Army

Research Institute for the Behavioral and Social Sciences

March 1986

Approved for public release; distribution unlimited.

TM 870 384

BEST COPY AVAILABLE

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Research accomplished under contract
for the Department of the Army

American Institutes for Research

Technical review by

Neil Schmitt (Michigan State University)
John J. Mellinger
William D. Sprenger (Shippensburg University)
Laurel Oliver
Paul van Rijn
Hilda Wing
William Haythorn

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARI Technical Report 603	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DEVELOPMENT OF OFFICER SELECTION BATTERY FORMS 3 AND 4		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report 10/80 - 9/83
		6. PERFORMING ORG. REPORT NUMBER AIR-86900-9/83-TR-1,2,3
7. AUTHOR(s) M. A. Fischl (ARI); Dorothy S. Edwards, John G. Claudy (AIR); and Michael G. Rumsey (ARI)		8. CONTRACT OR GRANT NUMBER(s) MDA-903-80-C-0701
9. PERFORMING ORGANIZATION NAME AND ADDRESS American Institutes for Research 1055 Thomas Jefferson Street, NW Washington, DC 20007		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263731A792
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333		12. REPORT DATE March 86 13. NUMBER OF PAGES 65
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

Technical quality of this research was monitored by M. A. Fischl

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Test development	Test fairness	Cognitive tests
Test standardization	Army ROTC	Noncognitive tests
Test validation	Officer selection	OCS selection

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes the development, standardization, and validation of two parallel forms of a test to be used for assessing young men and women applying to ROTC. Fairly extensive job analysis work established a content basis for initial test item development: 1,400 experimental items in 12 job relevant content areas were prepared and administered to 3,306 college juniors who were enrolled in the Advanced ROTC program. Professors of Military Science completed a Cadet Rating Form on each (continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued) (ARI Technical Report 603)
student covering Officer Potential and six scales relating to dimensions of Officer leadership.

The sample was stratified to conform to the 1980 national distribution of SAT scores and to consist of 18% black cadets, 5% other nonwhite cadets, and 10% female cadets. Item analyses were performed using both the Officer Potential rating and the sum of the Officer Dimension ratings as criteria. Separate analyses were also performed for gender and ethnic subgroups. Items that would yield the most valid tests with the least gender or ethnic impact were selected for the final forms.

For standardization, the tests were administered to college sophomores in military science courses, and the samples were again stratified to conform to the national distribution of SAT scores and the same gender and ethnic proportions. Results indicated that the tests were essentially equivalent, easily readable, and of high reliability. Separate norm tables were prepared for the two forms because of slight differences at the extremes of the distributions.

A small separate investigation involved administration of the test to a sample of high school seniors, since ROTC selection tests are sometimes used at this level for admission to military junior colleges or for scholarship purposes. As expected, scores for this group were lower. Since the high school sample was small and very likely nonrepresentative, the high school norms that were prepared were considered provisional.

Criterion-related validity was investigated twice. The test forms were administered to samples of senior ROTC cadets, and faculty ratings of leadership characteristics and officer potential were obtained for use as criteria. One of the test forms was also administered to Second Lieutenants in their first assignments (at Officer Basic Courses). Final course grades were obtained from the schools on these officers. Data analyses included correlations with criteria and, for the student sample, regression analyses for each form separately by gender and ethnic subgroups as well as the total group.

On the basis of the analyses performed, the Officer Selection Battery was concluded to be empirically and content valid and of comparable validity for ethnic and gender subgroups, with no indication of differential validity or regressions.

Technical Report 603

Development of Officer Selection Battery Forms 3 and 4

M. A. Fischl
Army Research Institute

Dorothy S. Edwards and John G. Claudy
American Institutes for Research

Michael G. Rumsey
Army Research Institute

Selection and Classification Technical Area
Newell Kent Eaton, Chief

Manpower and Personnel Research Laboratory
Joyce L. Shields, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

March 1986

Army Project Number
2Q263731A792

Manpower and Personnel

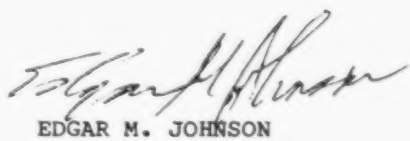
Approved for public release; distribution unlimited.

FOREWORD

This report describes the development, standardization, and validation of Forms 3 and 4 of the Officer Selection Battery (OSB), the Army's all-new test for assessing applicants to ROTC.

There are ROTC detachments at approximately 300 colleges and universities in the United States. These schools produce approximately three fourths of all commissioned officers. The Officer Selection Battery will be used in these schools as one part of a comprehensive assessment procedure intended to identify individuals who would serve well and satisfyingly as Army officers.

Forms 3 and 4 of the Officer Selection Battery are of job-relevant content, appropriate difficulty, high reliability, and state-of-the-art validity and fairness for minorities and women.



EDGAR M. JOHNSON
Technical Director

DEVELOPMENT OF OFFICER SELECTION BATTERY FORMS 3 AND 4

EXECUTIVE SUMMARY

Requirement:

To develop two equivalent forms of a paper-and-pencil, group administrable, test for use in assessing young men and women applying to Advanced Army ROTC.

Procedure:

Earlier research had performed an analysis of the Second Lieutenant's job resulting in a set of officer job dimensions. These dimensions were used to form the content basis on which test forms should be developed. Fourteen hundred test items in 12 job-relevant content areas were prepared for experimental tryout; they were administered to students in college Army ROTC programs just as they entered their junior year. Each student was rated by the ROTC faculty on scales describing officer leadership behavior and on an overall Officer Potential Scale.

The sample was stratified to conform to the 1980 national distribution of SAT scores and further stratified to contain 18% black representation, 5% other nonwhite representation, and 10% females. Item analyses were performed for the total sample and separately for gender and ethnic subgroups, and 110 items were selected for each of two forms of a test. These test forms were administered to 5,282 cadets during their sophomore year in Military Science II, when they would normally apply for Advanced ROTC, and the sample was stratified to reflect the same parameters as the item analysis sample.

Timing data indicated that virtually every cadet could complete the Officer Selection Battery (OSB) in available class periods, readability analyses indicated that the test is easily readable by persons with a ninth-grade education, and reliability coefficients were satisfactory.

The OSB was also administered to cadets during their senior year in Military Science IV, and faculty ratings were obtained. Separate correlational and regression analyses were performed for the total sample and for ethnic and gender subgroups; and one form was administered to Lieutenants in seven Officer Basic Courses. Correlation with final course grades was investigated.

Findings:

On the basis of the validity analyses performed, the OSB was concluded to be empirically and content valid, of comparable validity for ethnic and gender subgroups, with no indication of differential validity or regression.

Tables to convert raw test scores to Army Standard Scores were prepared, one for each form, and a separate investigation prepared provisional tables for use with high school seniors.

Utilization of Findings:

The OSB is used as part of the Precommissioning Assessment System in every ROTC program. Results from the OSB are used with other indicators as tools in the selection of Advanced ROTC cadets.

CONTENTS

	Page
INTRODUCTION	1
JOB ANALYSIS	1
TEST SPECIFICATIONS	2
Identification of Test Content Categories	2
Specifications for Item Construction	2
ITEM ANALYSIS DATA COLLECTION AND DATA PROCESSING	4
Assembly of Experimental Tests	4
Criterion Ratings	4
Other Test-Related Material	7
Shipment to Schools	7
Stratification of the Sample	8
Item Analysis	12
Timing Data	12
FINAL TEST FORMS	13
STANDARDIZATION DATA COLLECTION AND ANALYSES	13
Shipment of Materials	13
Initial Returns	13
Second Shipment	13
Stratification of the Sample	15
Interform Equivalence	15
Test Difficulty	16
Test Timing	16
Test Readability	17
Test Reliability	18
Norms, Raw to Army Standard Score Conversions	18
PERFORMANCE OF HIGH SCHOOL SENIORS	19
VALIDATION DATA COLLECTION AND RESULTS	21
ROTC	21
Officer Basic Courses	22
Validity in ROTC, Faculty Ratings as Criterion	23
Validity in Officer Basic Courses, Final Grades as Criterion	25
Test Fairness	26

CONTENTS (Continued)

Page

SUMMARY AND CONCLUSIONS 29

REFERENCES 32

APPENDIX A. DEFINITIONS OF IDENTIFIED DIMENSIONS OF OFFICER JOB PERFORMANCE A-1

B. TEST ITEM TYPES B-1

C. PLAN FOR ADDRESSING EEO FAIRNESS IN OSB C-1

D. WEIGHTING PROCEDURE D-1

E. READABILITY MEASURES E-1

F. COMPARISON OF OFFICER SELECTION BATTERY WITH TWO TESTS IN USE AS SELECTORS FOR ARMY OFFICER PRECOMMISSIONING TRAINING PROGRAMS F-1

LIST OF TABLES

Table 1. Officer job dimensions and test content 3

2. Target difficulty pattern 4

3. Distribution of items in OSB test booklet pairs 5

4. Number of complete sets of data by academic year, gender, and ethnic group 7

5. Initial target SAT distribution 9

6. Revised target SAT distribution 10

7. Target gender-by-ethnic-group distribution 10

8. Item analysis sample by SAT scores 11

9. Item analysis sample by gender and ethnic group 11

10. Time taken to complete each Officer Selection Battery booklet. 12

11. Item content of Officer Selection Battery 14

12. Distribution of item difficulty 14

LIST OF TABLES (continued)

Page

Table 13. Number of complete sets of MS-II data, by gender and ethnic group 15

14. Raw score means and standard deviations of Forms 3 and 4. 16

15. Percentage of students still working after 50 minutes . . . 17

16. Readability of Forms 3 and 4 using six different measures 18

17. OSB reliability estimates 19

18. OSB means and standard deviations in Army Standard Scores 20

19. Sample of ROTC detachments testing MS-IV cadets 21

20. MS-IV sample by gender and ethnic group 22

21. Officer Basic Course sample 23

22. Descriptive statistics for MS-IV sample 24

23. Correlation with rated Officer Potential and with sum of leadership dimension ratings 24

24. Correlation of OSB Form 3 with final grade in Officer Basic Courses 26

25. Validity of Officer Selection Battery by ethnic and gender subgroups 27

26. Regression coefficients for total samples and gender and ethnic subgroups 28

27. Statistical tests of difference in regression slopes between total samples and each subgroup 28

LIST OF FIGURES

Figure 1. Cadet Rating Form 6

2. Regression of criterion ratings on OSB scores for black and white samples and for total group--Form 3 30

3. Regression of criterion ratings on OSB scores for male and female samples and for total group--Form 3 30

LIST OF FIGURES (continued)

Page

Figure 4. Regression of criterion ratings on OSB scores for black and white samples and for total group--Form 4 30

5. Regression of criterion ratings on OSB for male and female samples and for total group--Form 4 30

INTRODUCTION

This report describes the development, standardization, and validation of an objective, group administrable, paper-and-pencil test for assessing men and women applying for Advanced Army ROTC officer training.

The test development effort was initiated in response to a determination by the Army that the procedures which had been in use for selecting students for officer precommissioning training were ready for replacement. The need was expressed for a paper-and-pencil test which would identify and qualify those individuals with a high probability of succeeding in the military community (Department of the Army, 1978).

JOB ANALYSIS

The first requirement in test development is a determination of the important elements of job success, what Guion (1976) has termed "criterion constructs." A recently completed job analysis effort (Rogers, Lilley, Wellins, Fischl, & Burke, 1982) constituted the primary source of information for this determination. Structured interviews were conducted with Lieutenants in combat arms, combat support, and combat service support branches to determine what activities were required of them in the performance of their jobs. This process began with the development of questions from available printed information about Army officer jobs, and other questions were directed at obtaining data on Lieutenant activities, problems, procedures, knowledges, and skills required on the job.

A second aspect of the job analysis involved the development of the constructs from critical incidents of successful or unsuccessful Lieutenant performance obtained during group interviews with Captains who supervised Lieutenants. Each of the two approaches identified performance dimensions, which were then administered to another 89 Captains who evaluated them for relevance and importance. Those relevant dimensions which met a predetermined criterion level of rated importance were retained and combined to form the following categories: communication, interpersonal manner, administration, decision making, initiative, and technical knowledge.

The categories above constituted a tentative list of the criterion constructs, which was compared with other recent literature describing dimensions of the Lieutenant's job. The literature reviewed included the work of Clement and Ayres (1976); Olmstead, Cleary, Lackey, and Salter (1973); Sitterson, Davis, and Korotkin (1974); Klemp, Munger, and Spencer (1977); Wellins, Rumsey, and Gilbert (1980); and specific descriptions of behaviors observed during a 3-day exercise in a simulated combat environment (Helme, Willemin, & Grafton, 1974).

The comparison of the tentative construct list with the other reported job analyses resulted in the addition of one new construct, labeled "combat performance." The final set of seven dimensions of officer job performance, with definition of each, appears as Appendix A.

TEST SPECIFICATIONS

Identification of Test Content Categories

The selection of test content involved identifying item types judged to be associated with the constructs and with promise for predicting officer performance. With respect to the dimensions of interpersonal manner, initiative, decision making, and administration, although it would not be possible in a paper-and-pencil test to actually demonstrate these behaviors, items using scenarios presenting problem situations with choices of solutions seemed to afford some potential for assessing characteristics of this nature. This approach guided the specification of scenarios depicting a variety of problems involving decision making, interpersonal performance, initiative, and administrative situations.

In the case of the communication dimension, two item types, which could be linked rationally to major components of the dimension, were specified. Verbal ability would reflect the individual's understanding of components of a communicated message, while certain aspects of problem solving would serve double duty as indicators of an individual's ability to combine the components in a logical manner.

Concerning the dimensions of technical knowledge and combat performance, which subsumed behaviors not easily definable, prior research by Helme, Willemin, and Grafton (1974) was consulted. Based on their research it was possible to specify quantitative ability, technical interests, and knowledge about physical sciences, history, politics, and culture as categories linked to technical knowledge; and ruggedness, stress tolerance, and knowledge of nature sports as categories linked to the combat dimension. The same research effort indicated that knowledge about tools, machines, and equipment was associated with both technical and combat performance. Finally, a set of spatial visualization items was specified for its judged relevance to map reading and land navigation aspects of combat performance.

An approximate correlation of officer job dimensions and test item types is shown in Table 1, and a more detailed presentation of specific kinds of items within each of the types appears as Appendix B.

Specifications for Item Construction

With item content specified, attention was next turned to administrative and statistical specifications for item construction. These specifications were developed in the context of certain general constraints established relative to the test. The test would be a four-alternative, multiple choice, power test. It was to be administrable during ROTC class, and no more than two 50-minute class periods could be devoted to testing. Two forms would be required, to allow for retest if necessary and to afford greater test security.

A wide-range difficulty pattern was specified for the test, and for each of the item types. Since Officer Candidate School applicants must have a minimum of 2 years of college, and applicants to the ROTC Advanced Course are usually college sophomores, the college sophomore population became the primary reference group for the test. Recognition was also given, though, to the

expectation that the test might be needed for some selection made from among high school seniors, and for some scholarship decisions. Hence the wide range of difficulty; but the primary target was the college sophomore, and the proportions of items prescribed for the various difficulty levels were maximum in the range of current and anticipated cutting scores where selection decisions would be made. Table 2 presents this pattern of item difficulty.

Table 1

Officer job dimensions and test content

Dimension	Type of test item
Administration	Verbal Problem solving General information
Communication	Verbal
Combat performance	Spatial visualization Assertiveness Initiative
Decision making	Problem solving
Interpersonal manner	Social problem solving
Technical knowledge	Quantitative Mechanical information General information

The issue of ethnic and gender fairness was a central one in this development. A comprehensive plan was developed to address test fairness at every stage of the development process. Key elements of this plan were the review of every test item by minority/female reviewers, attention to test instructions, and appropriate sampling of minorities and women for item analysis and standardization data collection. The full plan is in Appendix C.

In accordance with the specifications established, a total pool of 1,400 test items was assembled. Each item was subjected to a minimum of three reviews, two for technical content, accuracy, and conformance with good item-writing principles. The third review was conducted to look specifically for offensiveness or possible ethnic or sex bias in the items. This review was made by a senior-level researcher with experience in issues of ethnic and sex bias in a number of different areas in education and job performance, who was not involved in the item preparation, and is, herself, a member of an ethnic minority.

Table 2

Target difficulty pattern

Item difficulty[a]	% of items
.01-.06	3
.07-.15	6
.16-.30	10
.31-.49	14
.50-.68	20
.69-.83	20
.84-.92	14
.93-.97	10
.98-.99	3

[a]Percentage of nationally representative sample of college sophomores answering correctly.

ITEM ANALYSIS DATA COLLECTION AND DATA PROCESSING

Assembly of Experimental Tests

Early thinking had been toward a 170-item test, divided into two 85-item booklets to be administered in successive class periods. One form of the test, assembled this way, was tried out on some young people close to college age. The result of this try-out on eight individuals showed clearly that the 85-item format was too long. Accordingly, it was decided to reduce the test length to 100-110 items, and experimental booklets were so assembled for item analysis administration. No individual would be administered all 1,400 items. Reflecting the way the final test would be packaged, experimental analysis booklets were assembled in two 50-item pairs (e.g., 1 and 2, 3 and 4 ... 27 and 28) that would be administered in successive class periods at the various schools. The distribution of items within pairs of booklets is shown in Table 3.

Criterion Ratings

To estimate item validity, provision was made to obtain faculty appraisals of the overall "Officer Potential" of the cadet examinees, and appraisals of their observed performance on the six dimensions of Initiative, Decision Making, Administration, Communication, Interpersonal Manner, and Technical Knowledge. The rating form used for this purpose is shown as Figure 1.

An instruction manual was also provided to raters, which included descriptions of the five scale points and bench mark behavioral descriptions for anchor points on the individual dimensions.

Table 3

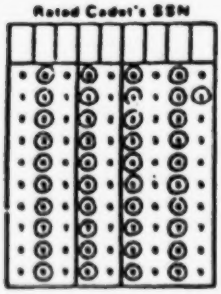
Distribution of items in OSB test booklet pairs

Category	No. Items	Booklet pairs 1&2	3&4	5&6	7&8	9 &10	11 &12	13 &14	15 &16	17 &18	19 &20	21 &22	23 &24	25 &26	27 &28
Verbal	260	25	10			25	20			30	10	35	20	55	30
Quantitative	200	25		30		25		30			30		30		30
Spatial: Map	40						20		20						
Cubes	20						10		10						
3-D	20									20					
Problem solving	199	25	19			15	15	20	25	15		45	10		10
General info.	265	25	30	30	40	15	15	30	35	15			20	10	
Tools	60									20		20		20	
Tech. interest	80		20		20				10		20				10
Assertiveness	60				10			20					20		10
Initiative	40		10	10	20										
Ruggedness	40						20				20				
Social comp.	59			19		20					20				
Managerial	57		11	11	10									15	10
	1,400														

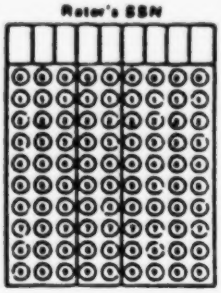
CADET RATING FORM

Please fill in the information requested below

Rated Cadet's SSN



Rater's SSN



Rated Cadet's Name (Print) | Institution (Print) | Rater's Name (Print)

INSTRUCTIONS

PART I: OFFICER POTENTIAL

Using the rating scale below, blacken one circle to indicate your best estimate of this cadet's potential as an Army officer. Include in your rating all of the abilities and qualities that you believe an officer needs, and indicate what kind of officer you think this cadet will become. In making your rating of potential, consider the cadet's demonstrated performance relative to the performance of other ROTC cadets.

Bottom 10% (Marginal)	Next higher 20%	Middle 40% (Average)	Next higher 20%	Top 10% (Outstanding)
①	②	③	④	⑤

PART II: OFFICER DIMENSIONS

The dimensions you are to rate the cadet on in this part are briefly described below. To rate the cadet, use the behavior descriptions shown in the accompanying "Instruction Manual for Using Cadet Rating Form." Compare this cadet's demonstrated performance with the descriptions in that manual, and blacken the number on each dimension scale below that you think best fits the cadet.

Dimension	Rating Level: Marginal		Average		Outstanding
1. Initiative. Active attempts to achieve goals; self-starting rather than passive acceptance	①	②	③	④	⑤
2. Decision making. Demonstration of problem analysis, judgment and decisiveness in response to problems	①	②	③	④	⑤
3. Administrative skills. Effectiveness in planning, organizing, delegation and administrative control	①	②	③	④	⑤
4. Communication. Clarity and effectiveness of expression of ideas or desires in writing and orally (formally and informally)	①	②	③	④	⑤
5. Interpersonal skills. Utilization of appropriate styles and methods of influence in guiding others toward task accomplishment	①	②	③	④	⑤
6. Technical skills. Level of understanding and ability to use technical/professional information	①	②	③	④	⑤

Figure 1. Cadet Rating Form.

Other Test-Related Material

A manual for administration was prepared, as was an Examiner Report Form. This form was intended to be used at each testing session to report the number of examinees, the number still working at 40 and 50 minutes after the start of the test, and any irregularities in the testing that might affect inclusion of an answer sheet in the analysis; e.g., if a student became ill and left after completing only part of the test.

Two optically scannable answer sheets, one for use with each of the paired booklets, were provided for each examinee. These forms requested social security number, for collating purposes, and gender and ethnic group information.

Finally, a variety of administrative instructions and guidelines was prepared.

Shipment to Schools

Test material was shipped to every ROTC host institution (276) in sufficient quantity to test every student in the college or university who was enrolled in Military Science III (MS-III).

Complete data--two answer sheets and a rating scale--were received from 8,778 examinees at 233 schools. Table 4 shows a distribution of the returned data by important moderators.

Table 4

Number of complete sets of data by academic year, gender, and ethnic group

Academic year	N	Gender	N	Ethnic group	N
Freshman	1,380	Male	7,373	White	6,227
Sophomore	1,078	Female	1,397	Black	1,676
Junior	3,306			Hispanic	541
Senior	154			Indian	27
Graduate	45			Asian	169
				Other	89
Missing academic year	2,815	Missing gender	8	Missing ethnic group	49
Total	8,778		8,778		8,778

The full set of 8,778 returns was reduced by limiting the item analyses to the responses of the 3,306 academic juniors. The intention of the test development was to construct an instrument applicable to the majority of applicants to Army precommissioning training programs, and these individuals typically have 2 years of college. Inasmuch as this data collection took place in early autumn of the school year, the academic junior had just completed 2 years of college and was considered most representative of eventual program applicants.

Stratification of the Sample

Precise specification of sample characteristics was considered important, to allow meaningful interpretation of such statistics as item difficulty, and to permit successive generations of the test to be constructed to the same parameters as Forms 3 and 4 herein developed and thus be additional equivalent forms.

The sample of academic juniors was weighted to conform to the 1980 national distribution of College Entrance Examination Board Scholastic Aptitude Test (SAT) total scores: Mean = 890, Standard Deviation = 206. SAT scores or their equivalent values from the American College Testing Program or the Army's Cadet Evaluation Battery had been requested of all examinees, but were only available on 2,805 of the 3,306 academic juniors. This sample was also weighted to reflect appropriate ethnic and gender representation. Since census data indicate that black youths comprise approximately 13 to 15% of the 17- to 24-year-old population, the sample was designed to contain no fewer than 15% black youths. In addition, since the Army's accession plans will provide for a female strength of approximately 10% of the force, this percentage was taken as the guiding target for the female portion of the sample.

The procedure used to weight the sample is described in detail in the next few pages, as a matter of record, since it was used again in connection with the development of norms and is intended as a reference for developing future alternate forms of the OSB.

SAT Distribution. A distribution of SAT Total scores which is normal in shape, has a mean of 890, and a standard deviation of 206, can be divided into 25 intervals containing the expected percentages shown in Table 5.[1] A general rule was adopted to permit no interval to contain fewer than one-half of 1% of the cases. With 2,805 cases, i.e., the number on whom SAT equivalent scores were available, one-half of 1% equals 14 cases. The actual distribution of the 2,805 cases had six intervals, at the tails, with half or fewer of this required frequency. Accordingly, the top four intervals were combined to yield a frequency of 16, and the bottom two intervals were combined to yield a frequency of 12, slightly less than desired, but addition of the next interval would have overweighted the first interval threefold. The corresponding expected

---

[1]Table 5's distribution of expected percentages departs slightly from symmetry because the mean does not fall at the mid-point of an interval, and because it is closer to one of the limiting end points than to the other. The interval size and end points of the distribution were selected on convenient a priori grounds.

percentages were also combined from the target distribution, resulting in a revised target distribution consisting of 21 weightable intervals, as shown in Table 6.

Table 5

Initial target SAT distribution

SAT total interval	Expected percentage in interval
400-420	1.19
430-470	1.00
480-520	1.63
530-570	2.49
580-620	3.61
630-670	4.92
680-720	6.32
730-770	7.68
780-820	8.76
830-870	9.49
880-920	9.66
930-970	9.27
980-1020	8.36
1030-1070	7.16
1080-1120	5.77
1130-1170	4.36
1180-1220	3.14
1230-1270	2.11
1280-1320	1.35
1330-1370	.81
1380-1420	.45
1430-1470	.24
1480-1520	.13
1530-1570	.05
1580-1600	.05
Total	100.00

Gender and Ethnic Group Distribution. The 2,805 usable cases were categorized into one of five gender-by-ethnic-group categories with the target distribution as shown in Table 7.

Calculation of Sample Weights. The remaining steps were those mathematical ones necessary to bring the observed distribution of 2,805 academic juniors into congruence with the target SAT and gender-by-ethnic-group distributions. The process is described in detail in Appendix D; the result is shown in Tables 8 and 9.

Table 6

Revised target SAT distribution

Interval number	SAT score interval	Expected percentage
1	400-470	2.19
2	480-520	1.63
3	530-570	2.49
4	580-620	3.61
5	630-670	4.92
6	680-720	6.32
7	730-770	7.68
8	780-820	8.76
9	830-870	9.49
10	880-920	9.66
11	930-970	9.27
12	980-1020	8.36
13	1030-1070	7.16
14	1080-1120	5.77
15	1130-1170	4.36
16	1180-1220	3.14
17	1230-1270	2.11
18	1280-1320	1.35
19	1330-1370	.81
20	1380-1420	.45
21	1430-1600	.47
Total		100.00

Table 7

Target gender-by-ethnic-group distribution

Category	Target %
1. White males	69.0
2. Black males	16.0
3. Other males	4.5
4. White females	8.0
5. Black and other females[a]	2.5
Total	100.0

[a]This category consists of 2% black females and 0.5% other females.

Table 8

Item analysis sample by SAT scores

SAT total interval	Observed frequency	Weighted frequency
400-470	12	876
480-520	30	652
530-570	33	996
580-620	40	1,444
630-670	143	1,968
680-720	188	2,528
730-770	167	3,072
780-820	234	3,504
830-870	214	3,796
880-920	210	3,864
930-970	319	3,708
980-1020	266	3,344
1030-1070	241	2,864
1080-1120	180	2,308
1130-1170	152	1,744
1180-1220	121	1,256
1230-1270	85	844
1280-1320	79	540
1330-1370	47	324
1380-1420	28	180
1430-1600	16	188

Table 9

Item analysis sample by gender and ethnic group

Gender/ethnic group	Observed frequency	Weighted frequency
Male		
White	1,929	27,600
Black	286	6,400
Other	137	1,800
Female		
White	303	3,200
Black and other	150	1,000
Total	2,805	40,000
White	2,232	30,800
Black	420	7,200
Other	153	2,000

Item Analysis

The Officer Potential scale of the Cadet Rating Form was utilized as the primary external criterion for the item analysis. In addition, as an internal criterion, each item was correlated with a score (percentage correct) on the set of all items of its type. The internal analysis was performed for the total sample only, while correlations with the Officer Potential rating were computed for all males, all females, all blacks, and all whites, in addition to the total sample. No separate item analysis was performed of the Hispanic group inasmuch as this group was very small and most of the cases came from a single large Puerto Rican university.

For each option of each item, the raw and weighted N selecting the option, the percentage selecting it, the criterion mean and standard deviation, the point biserial correlation, and the Brogden-Clemans biserial were calculated. These statistics were utilized in making item selections for the final forms of the test. Selections were made so as to maximize item external validity consistent with satisfying specifications for item content and difficulty, and to minimize black-white, male-female differences in item external validity and difficulty.

Timing Data

During administration of the item pools, assembled into 50-item booklets, information was obtained on how much time these booklets required to complete. This is presented in Table 10, a report of the percentage of institutions in which examinees completed a booklet within certain time intervals.

Table 10

Time taken to complete each Officer Selection Battery booklet

All examinees completed in:	% of schools	Cumulative %
30 minutes or less	20	20
31-35 minutes	26	46
36-40 minutes	30	76
41-45 minutes	9	85
46-50 minutes	10	95
More than 50 minutes	5	100

The 50-minute class period appeared sufficient for completing the 50-item test booklet and, since nearly half the examinees had finished in 35 minutes or less, it was decided that the inclusion of 55 items in each of the OSB test booklets would not be excessive.

FINAL TEST FORMS

Utilizing the item analysis rules and the outcomes of the timing investigations, a final selection of 110 items for ech of two forms was made and packaged to contain half of its items in each of two 55-item test booklets. The item content of the final test forms is shown in Table 11, and the distribution of item difficulty values is presented in Table 12.

STANDARDIZATION DATA COLLECTION AND ANALYSES

Shipment of Materials

Test material for the standardization data collection was shipped to 276 ROTC detachments in March 1982. Inasmuch as the intended operational use of the test was to make selections to Advanced ROTC from among Military Science II (MS-II) students, sufficient quantity was provided to test every student enrolled in MS-II. The closing enrollment figures for 1981 were used to estimate the size of the shipment, augmented by a slight overage. The shipment consisted of test booklets, optically scannable answer sheets, an administration manual, an Examiner Report Form for use in obtaining timing information and reporting any irregularities, administrative instructions, and data collection guidelines.

Initial Returns

Shortly after material was sent to the schools it became clear that there would be heavy attrition in the sample. Some schools had already closed for the summer. For certain others, ROTC classes had either completed their year or were so tightly scheduled that the two class periods required for the test were simply not available. Finally, at some schools that were still open, final examinations were in progress or imminent, and students declined to take the OSB so they could study. This was permissible, of course, in accordance with the Privacy Act and its statement that appeared on the back of each test, pointing out the voluntary nature of the task.

Complete data on academic sophomore MS-II students were available from only 164 schools, approximately 60%, yielding 2,714 cases. Examination of these returns showed clearly that the 40% of the schools not responding was nonrandom and the possibility of thus inadvertently biasing the sample of schools and students responding could not be overlooked. In addition, since half of the students were administered Form 3, the other half Form 4, the number of cases would be too small for analysis of other than the total sample, so any differences between or among gender and ethnic groups could not be examined. It was therefore decided that a second shipment of material would be required in order to augment the sample.

Second Shipment

Schools that had either not responded or had returned data on fewer than 50% of their academically aligned MS-II students in the spring 1982 class were requested to test the fall 1982 class of MS-II students. There were 161 such schools.

Table 11

Item content of officer selection battery

Type of item	Number of items	
	Form 3	Form 4
Verbal	32	32
Quantitative	25	25
General information	20	20
Problem solving		
General problems	18	15
Assertiveness problems	2	2
Initiative problems	2	3
Managerial problems	2	3
Social problems	1	2
Spatial		
Folding/unfolding geometric forms		4
Map reading	4	4
Three-dimensional figures	4	
Total	110	110

Table 12

Distribution of item difficulty

Item difficulty	Number of items	
	Form 3	Form 4
0.01-0.06	1	0
0.07-0.15	0	1
0.16-0.30	6	5
0.31-0.49	12	12
0.50-0.68	31	25
0.69-0.83	34	39
0.84-0.92	14	16
0.93-0.97	10	10
0.98-0.99	2	2

Data were returned from 159 schools; two schools had unusual problems that prohibited their participation. The data were combined with those of the spring 1982 administration. The number of complete sets of data on academically aligned MS-II students is shown in Table 13.

Table 13

Number of complete sets of MS-II data, by gender and ethnic group

Group	Form 3	Form 4
Males	2,259	1,978
Females	577	468
Black	589	489
Hispanic	133	61
White	2,066	1,845
Other	48	51
Total	2,836	2,446

Stratification of the Sample

College Entrance Examination Board Scholastic Aptitude Test (SAT) scores or their equivalent, and demographic information, had been requested of all examinees. These data were used to stratify the sample to the same parameters as for item analysis.

Interform Equivalence

Table 14 shows the raw score means and standard deviations of the two forms of the test, for the total sample and for gender and ethnic subgroups. The two forms of the test have highly similar statistics, and the goal of producing two equivalent forms seems to have been met. Note is taken that the Form 3 and Form 4 means are essentially within rounding of one another, with the respective standard deviations about one raw score point apart.

The similarity of the two test forms is equally manifest in the gender subgroups and in the white ethnic group. The slightly larger mean difference between the forms in the black ethnic group can be attributed to sampling variation, because the SAT mean of the black cadets administered Form 3 turned out to be 13 points higher than that of the black cadets administered Form 4. It is also suggested that interform differences in the Hispanic ethnic group be ignored inasmuch as the size of the two samples from this group was small.

Table 14

Raw score means and standard deviations from Forms 3 and 4

Group		Form 3	Form 4
Total stratified sample:	Mean	74.82	74.40
	S.D.	14.27	15.39
Black sample:	Mean	59.19	56.25
	S.D.	12.59	13.82
White sample:	Mean	79.24	79.14
	S.D.	11.23	11.97
Hispanic sample:	Mean	62.27	67.31
	S.D.	14.93	15.34
Male sample:	Mean	75.08	74.68
	S.D.	14.13	15.38
Female sample:	Mean	72.63	72.05
	S.D.	15.18	15.31

Test Difficulty

The raw score means shown in Table 14 define the average difficulty of the 110-item test, which is thus 68% for each of the forms. Compared to the total stratified sample (i.e., cadets in general), the test is not appreciably harder for women nor appreciably easier for white examinees. The test is more difficult for black examinees, their mean being approximately a standard deviation lower than the mean of the total sample. That outcome, although disappointing in view of the effort expended to minimize group differences, is consistent with what is more commonly reported in the literature (e.g., Wigdor & Garner, 1982). Nevertheless, it should be recognized that the test will have an adverse impact on many black examinees. This means that issues of test validity, indications of presence or absence of differential validity, and appropriateness of a single regression line become very important. These are dealt with later in the report.

Test Timing

Table 15 presents timing information. The test is intended to be a power test, administrable in two 50-minute class periods. It was thus important to determine what percentage of examinees completed each of the 55-item half-tests in the 50-minute class period. The testing in the schools was timed and, as may be seen, the goal of producing test modules that nearly all students could finish in a single class period seems to have been accomplished.

Table 15

Percentage of students still working after 50 minutes

	Form 3 N = 4,221[a]	Form 4 N = 3,731[a]
Booklet 1	0.7	1.4
Booklet 2	0.6	1.3

[a]All MS-II cadets, irrespective of academic class or any missing non-OSB data.

By usual standards of 95% completion, the OSB can be considered to be of appropriate length and minimal speed loading.

Test Readability

The reading comprehension level that a test such as the OSB requires is difficult to evaluate, because tests typically include types of items that do not lend themselves to the available reading measurement procedures. Most reading measures depend upon such variables as word length or sentence length, which are not applicable to vocabulary, mathematics, and some general information items. For these, the response alternatives, and sometimes the question stems, consist of only one or two words. In addition, of course, many spatial items are totally nonverbal.

In order to obtain some evidence of readability, several formulas were applied to a selection of items from each of the test booklets. The items covered the more "densely packed" reading material, primarily from the problem solving content areas. This approach afforded a set of passages with the type of prose which readability formulas were designed to evaluate. If the approach errs at all, it is in the direction of overestimating the true reading demands of the test. That is, the test is probably easier to read than the results indicate, because of the large number of nonreading items.

The following readability formulas were applied:

Flesch Reading Ease
Sticht
FORCAST
Gunning Fog Index
Kincaid/Flesch
Flesch (Original)

A brief description of each has been placed in Appendix E.

Table 16 shows the readability scores of the OSB on the six measures. Two points seem apparent. First, the test is easy to read. The "worst case" measure of grade level, indicates about a ninth-grade--for a test designed to examine college sophomores. Thus, performance on the OSB should be expected to be a function of aptitude and knowledge, not reading skill. Second, the two forms of the test appear to be quite comparable in their reading demands. The small differences that are seen would be unlikely to affect test scores attained by a college population.

Table 16

Readability of Forms 3 and 4 using six different measures

Readability formula	Form 3	Level	Form 4
Flesch	73.9	Fairly easy/easy	80.2
Sticht	73.0	1-syllable words	71.0
FORCAST	9.1	Grade level	9.3
Gunning Fog Index	8.3	Grade level	8.8
Kincaid/Flesch	74.0	Fairly easy	80.0
Flesch (Original)	74.0	Fairly easy	80.0

Test Reliability

The only, even approximate, indication of test reliability available from a single test administration is internal consistency. Note should be taken that OSB was constructed to contain some fairly diverse content (Table 11). Internal consistency estimates for the two forms of the OSB were determined by computation of coefficient alpha, and more fundamental reliability was estimated by the standard error of measurement. Table 17 shows the results.

Internal consistency of the OSB is high, even though much of the content is intentionally heterogeneous. Further, for a 110-item test, standard errors of the order 4 raw score points can be considered to describe a very appropriate level of measurement precision.

Norms, Raw to Army Standard Score Conversions

Tables of equivalents were prepared, indicating the Army Standard Score equivalent of each raw score. The Army Standard Score scale has a mean of 100 and a standard deviation of 20. The procedure that was used to derive the equivalents was a smoothing of the weighted frequency distributions, using a seven-point weighted moving average procedure. This procedure leaves the mean and standard deviation intact, as well as the general shape of the distribution, but smooths the high and low points. A total of 10 iterations was applied to each of the distributions. Then, using the cumulative percent of cases for

each raw score value, a standard score was assigned corresponding to the percent of cases in a normal distribution which would fall below this raw score value, plus half the cases at that value. This procedure forces the standard score distribution to have a normal distribution shape.

Table 17

OSB reliability estimates

Estimate	Form 3 N = 4,221[a]	Form 4 N = 3,731[a]
Coefficient alpha	0.92	0.94
Standard error of measurement[b]	4.04[c]	3.77[c]

[a]For reliability estimation purposes the test scores of all MS-II cadets were utilized, irrespective of academic year or any missing non-OSB data.

[b]Calculated utilizing the coefficient alpha values as the reliability estimates.

[c]The corresponding Army Standard Score values of these standard errors are 5.66 for Form 3 and 4.90 for Form 4.

Although it would have been administratively preferable to end up with a single conversion table for both forms, Forms 3 and 4 cannot use the same conversion table. The means and standard deviations are virtually identical; but there are small differences at the extremes of the distributions, the smoothing out of which would have introduced an unacceptable degree of error. Thus, a separate conversion table was prepared for each of the two OSB forms.

For the reader's convenience, Table 18 provides in Army Standard Score units the comparison information to Table 14. The interform similarity in means and standard deviations is quite apparent, as is the one standard deviation difference between the means of the black sample and the total sample.

PERFORMANCE OF HIGH SCHOOL SENIORS

Although most selection decisions for officer precommissioning training programs are made from among applicants with 2 years of college, there are certain circumstances in which applicants are high school seniors. For example, there are six military junior colleges in the United States; ROTC offers a limited number of scholarships, applicants for which are high school seniors; the Military Academy accepts applications from high school seniors; in mobilization it is possible that the educational requirements for some officer training programs might be reduced. Given the college sophomore population as the primary standardization group for the OSB, but given the potential for use with a high school senior group, it seemed appropriate to investigate the test's generalizability.

Table 18

OSB means and standard deviations in Army standard scores

Group		Form 3	Form 4
Total stratified sample:	Mean	100.78	100.77
	S.D.	20.27	20.29
Black sample:	Mean	80.18	79.11
	S.D.	14.92	15.37
White sample:	Mean	106.60	106.48
	S.D.	17.55	17.51
Hispanic sample:	Mean	84.31	91.15
	S.D.	18.59	18.14
Male sample:	Mean	101.16	101.14
	S.D.	20.17	20.36
Female sample:	Mean	97.61	97.57
	S.D.	20.84	19.39

Generalizability to a high school group could not be assumed, in view of maturational/developmental changes expected to take place between ages 18 and 20, and in view of the two additional years of educational experience. Hence a sample of high school seniors was drawn from the Junior ROTC programs of 16 high schools in the South and Midwest sections of the United States. The intention here was not necessarily to draw a tightly representative sample from which conclusive inferences could be drawn, but to make an initial determination of the relevance of the OSB for high school seniors and to estimate whether score differences of any meaningful magnitude would be observed.

The high school samples, one administered Form 3, one Form 4, were stratified to conform to the 1980 SAT reference distribution, and summary statistics were calculated. Then the two high school samples were combined and, utilizing the same procedure as was utilized with the college sophomore samples, a norm table showing the Army Standard Score equivalent to each raw test score was prepared.

Results of the high school investigation should be considered tentative. In fact, most of the complete data for the investigation came from only one or two large schools. Nevertheless, the basic concern of the investigation was corroborated--the OSB is a more difficult test for high school seniors than it is for college sophomores. In these stratified samples the high school means were between five and six raw score points lower than the college means, although there was no appreciable difference in the standard deviations.

This finding suggests that if used with high school seniors, either a separate norm conversion table should be employed or a lower qualifying score required. The conversion table which has been prepared should be considered tentative, pending acquisition of data from a larger and possibly more representative sample of high school seniors.

VALIDATION DATA COLLECTION AND RESULTS

ROTC

At the time of standardization data collection among MS-II cadets, the ROTC faculty at a sample of 74 of these institutions was instructed to test their MS-IV cadets and to render an evaluation of their officer potential and their leadership characteristics on the same rating form that was utilized in the item analysis phase. This sample of schools is described in Table 19.

Table 19

Sample of ROTC detachments testing MS-IV cadets

Population characteristic	No. of schools in sample
ROTC region	
First	26[a]
Second	21
Third	14
Fourth	13
School ownership	
Public	49
Private	25
Historically black colleges	21
School size	
Large	18
Medium	37
Small	19

[a]This number includes two schools on the island of Puerto Rico.

An instruction manual provided raters with descriptions of the five rating levels and, for the individual leadership dimension ratings, benchmark behavioral descriptions for three anchor rating levels.

In addition to the tests, rating forms, and rating instruction manuals, each shipment to schools included an Administration Manual, an Examiner Report Form for reporting any irregularities, and administrative instructions and data collection guidelines. This material was shipped to the schools in March 1982.

Initial Returns. For the reasons described in connection with the standardization data collection, there was heavy attrition in the sample. Data were received from only 51 of the 74 schools, and examination of the returns showed clearly that the attrition was nonrandom. The possibility of inadvertent bias in the sample could not be overlooked. It was therefore decided that a second testing would be required in order to augment the sample.

Second Shipment. The second shipment of material to schools took place in November 1982. MS-IV testing was requested at the 23 schools that had not responded to the earlier request and at one additional school that had a substantial black enrollment (although not one of the historically black colleges), selected to ensure adequate black representation in the sample.

Data Available. Data returned from the schools of the second testing were combined with those received earlier, and Table 20 describes this sample. It should be noted that the Hispanic sample is small, particularly for Form 3. Moreover, most of the Form 4 results were from a single large institution on the island of Puerto Rico. Because of the small size and extreme geographic bias in the Hispanic sample, no satisfactory basis was obtained for estimating OSB validity among Hispanic cadets.

Table 20

MS-IV sample by gender and ethnic group

	Form 3	Form 4
Male	502	686
Female	75	102
Black	156	101
Hispanic	22	79
White	385	590
Not reported	14	18
Total sample	577	788

Officer Basic Courses

Upon being commissioned, the Lieutenant's first assignment is to an Officer Basic Course (OBC) in which he or she is instructed in the specific content of his or her officer branch and specialty. Since an implicit goal of a precommissioning selection instrument is to forecast early postcommissioning performance,

correspondence of OSB scores with OBC performance was investigated in a sample of such schools.

At the time of this research there were 14 OBCs. One of the test forms (Form 3) was administered to 577 Lieutenants in the following schools: Engineer, Field Artillery, Infantry, Military Police, Ordnance, Quartermaster, and Signal. The breakdown of number of examinees by school is presented in Table 21. Final course grades, in addition to test scores, were obtained for each examinee.

Table 21

Officer Basic Course sample

Course	N
Engineer	64
Field Artillery	267
Infantry	91
Military Police	59
Ordnance	31
Quartermaster	28
Signal	37
Total	577

Validity in ROTC, Faculty Ratings as Criterion

Table 22 presents the means and standard deviations for the MS-IV samples administered each of the two forms of the test.

Validity estimates were computed independently utilizing the global Officer Potential rating, and the sum of the individual leadership dimension ratings. These correlations are shown in Table 23.

The validity coefficients shown in Table 23 were calculated by pooling all examinees who had been administered a specified test form, Ns as shown in Table 22. This procedure permitted utilizing all of the data, but confounded any existing interschool differences in rating standards. An alternative was to calculate the coefficients separately by school and average them. Some 19 schools had tested fewer than 10 students each, and these were removed from the analysis sample. In each of the remaining 55 schools, correlations of OSB scores with the Officer Potential and sum of the individual leadership dimension ratings were calculated; then weighted mean validity coefficients were derived utilizing Fisher's Z transformation. The resulting counterpart values to those reported in Table 23 were 0.302 and 0.325 for Form 3, 0.342 and 0.324 for Form 4.

Table 22

Descriptive statistics for MS-IV sample

Variable	N	Mean[a]	Standard deviation[a]
Form 3	577	77.29	14.60
Rating on Officer Potential		3.55	1.02
Ratings on Officer Leadership dimensions:			
Initiative		3.59	1.12
Decision making		3.62	1.03
Administrative skills		3.59	1.02
Communication		3.67	1.04
Interpersonal skills		3.59	1.01
Technical skills		3.75	0.97
Sum of dimension ratings		21.81	5.55
Form 4	788	79.00	13.68
Rating on Officer Potential		3.58	1.02
Ratings on Officer Leadership dimensions:			
Initiative		3.67	1.06
Decision making		3.66	0.99
Administrative skills		3.60	1.02
Communication		3.68	1.00
Interpersonal skills		3.66	1.02
Technical skills		3.77	1.00
Sum of dimension ratings		22.05	5.39

[a]Test means and standard deviations are raw scores.

Table 23

Correlation with rated Officer Potential and with sum of leadership dimension ratings

	Form 3	Form 4
Officer Potential	0.205	0.285
Sum of dimension ratings	0.258	0.275

Inasmuch as any unreliability in the rating criterion would serve to attenuate predictor-criterion relationships, two approximate estimates of the reliability of the ratings were developed. First, the six individual leadership dimensions were treated the way items of a test would be, and an internal consistency coefficient, coefficient alpha, was calculated. Second, the global Officer Potential rating and the sum of the ratings on individual dimensions were treated the way scores on alternate forms of the same test would be, and a coefficient of equivalence was calculated. These values were 0.95 and 0.92 respectively. When they were utilized to correct the validity coefficients of Table 23, the corrected coefficients were 0.214 and 0.263 for Form 3, and 0.297 and 0.279 for Form 4. It should be noted that the most appropriate estimate of criterion reliability, not available, would have been interrater agreement, perhaps with an intervening period of time between the two ratings. Such a value would undoubtedly have been much lower than the two estimates obtained; hence the corrections applied are maximally conservative and the resulting coefficients should be viewed as underestimates of the true validity of the test forms.

An attempt also was made to determine if the process of selection from MS-II into the ROTC Advanced Course (MS III-IV) might have operated to restrict the range of ability present in the MS-IV sample of examinees. To the extent that this might have occurred, i.e., truncation of the bottom portion of the ability distribution, it would also have operated to attenuate predictor-criterion relationships observed in the restricted sample.

To assess the extent of any restriction, the OSB standard deviations of the MS-II samples utilized in norming the test forms were compared with their counterpart values in the MS-IV samples. The MS-II standard deviations for the Form 3 and Form 4 samples were 14.3 and 15.4 raw score points, and their MS-IV counterparts were 14.6 and 13.7. The OSB standard deviations were compared because that is the only test that the entire MS-II sample took. However, statistically OSB is a variable of incidental selection rather than the variable of explicit selection. Explicit selections into Advanced Course ROTC were made on College Board SAT scores for some cadets, American College Testing Program (ACT) scores for others, and a subscore from the Army's Cadet Evaluation Battery (CEB) for still others. Since many more cadets were admitted on CEB scores than on either of the other tests, CEB standard deviations were also compared. The MS-II CEB standard deviations for the Form 3 and Form 4 samples were 17.2 and 16.1 Army Standard Score points, and their MS-IV counterparts were 17.3 and 15.5. Apparently the CEB qualifying score was a minimum, which most applicants exceeded. In these samples range restriction does not appear to have been a factor, so no further correction of validity estimates was made.

Validity in Officer Basic Courses, Final Grades as Criterion

Table 24 presents the validity of the OSB in the sample of Officer Basic Courses investigated. These validity coefficients are substantially higher than those which utilized ROTC faculty ratings as criteria. This is not an uncommon event; ratings tend to result in lower estimates of test validity than do other criteria. Regardless, the level of OSB concurrence with OBC performance as shown in Table 24 is very substantial.

Table 24

Correlation of OSB Form 3 with final grade in Officer Basic courses

Course	N	Correlation coefficient
Signal	37	0.77
Quartermaster	28	0.64
Ordnance	31	0.58
Infantry	91	0.53
Military Police	59	0.52
Engineer	64	0.50
Field Artillery	267	0.45
Average[a]	577	0.52

[a]The averaging procedure utilized the r to Z transformation.

Test Fairness

An additional concern is the fairness of the OSB to women and to ethnic minorities. Accordingly, a number of analyses were performed to determine comparability of validity coefficients and regression lines for the different gender and ethnic groups as well as the total group.

The validity of the OSB for the various groups, against ratings in MS-IV ROTC, is shown in Table 25. What is very clear from this table is that the OSB is no less valid for women and ethnic minorities than it is for the total group or for whites and males separately. In fact, where there is any observed difference in validity coefficients, the value shown is actually higher in every case[2] for the group protected under the law than for the majority.

The sum of the leadership dimension ratings, the slightly more reliable of the two criteria, was regressed on the OSB for the total sample and for the ethnic and gender subsamples. The results are shown in Table 26. The similarity in these regression coefficients is striking. They appear to be tightly distributed around some single common value, except for one sample. The regression coefficients developed for the sample of black cadets administered Form 4, for unknown reasons, does not appear to fit the pattern.

[2]The one exception to this statement is the Hispanic sample administered Form 3. Because of the extremely small number of cases, it cannot be considered appropriately representative.

Table 25

Validity of Officer Selection Battery by ethnic and gender subgroups

Group	N[a]	Rated Officer Potential	Sum of dimension ratings
Form 3			
Black	156	0.21	0.27
White	385	0.20	0.24
Hispanic	22	0.10	0.33
Female	75	0.28	0.30
Male	502	0.19	0.26
Total sample	577	0.20	0.26
Form 4			
Black	101	0.39	0.34
White	590	0.26	0.23
Hispanic	79	0.26	0.31
Female	102	0.32	0.33
Male	686	0.27	0.25
Total sample	788	0.28	0.28

[a]Some examinees omitted ethnic group.

As a check on the observational analysis of the previous paragraph, t-tests were performed on the difference between the slope of the total sample's regression line, and the slopes of each of the gender and ethnic subgroups. The results are presented in Table 27, in which it is apparent that, by customary statistical significance criteria, the slope of no separate regression line is different from the slope of the line based on the total group. Plots of these regression lines are shown in Figures 2, 3, 4, and 5.

What then may be inferred concerning the gender and ethnic fairness of the OSB? In the standardization of the test it was observed that it was a more difficult examination for black applicants, on average, than for majority applicants. It was pointed out that such will probably result in an adverse impact on black examinees. Adverse impact, although undesirable, does not make a test unfair; it is, however, a signal that other properties of the test must be investigated. That has been done, with results indicating that the OSB has demonstrably nonzero criterion-related validity for all ethnic subgroups

Table 26

Regression coefficients for total samples and gender and ethnic subgroups

Form	Intercept	Slope
Form 3		
Total sample	15.14	0.06
Male	14.47	0.07
Female	16.21	0.06
Black	14.26	0.08
White	13.34	0.07
Form 4		
Total sample	14.53	0.07
Male	14.55	0.07
Female	13.80	0.09
Black	8.44	0.14
White	14.56	0.07

Table 27

Statistical tests of difference in regression slopes between total samples and each subgroup

Test	Difference in slopes[a]
Form 3	
Total--Male	0.01
Total--Female	0.00
Total--Black	0.02
Total--White	0.01
Form 4	
Total--Male	0.00
Total--Female	0.02
Total--Black	0.07
Total--White	0.00

[a]None of these differences is statistically reliable, $p > 0.05$.

examined, and that no evidence of differential validity was observed, i.e., the test is no less valid for minorities and women than for majority groups. Further, one regression line, based on the total sample, is appropriate for use with all groups studied.[3] By these analyses, and subject to cautions which have been introduced, the Officer Selection Battery may be considered to be race and gender fair.

SUMMARY AND CONCLUSIONS

This report describes the development, standardization, and validation of parallel forms of a 2-hour, group administrable, paper-and-pencil test for selecting men and women for Army officer training programs. A job analysis identified key dimensions of the Army Lieutenant's job, which formed the basis for the content of the instrument. A pool of 1,400 test items was prepared, in accordance with specifications for that content as well as specifications for item difficulty and test fairness. This pool was administered to ROTC cadets in all host institutions in the fall of 1981. Faculty evaluations of each tested cadet's officer potential comprised the external criterion against which the validity of each test item was evaluated, in samples stratified to conform to: (a) the 1980 national distribution of College Entrance Examination Board Scholastic Aptitude Test (SAT) scores; (b) an ethnic group composition of 18% black cadets, 77% white cadets, 5% cadets of other groups; (c) 10% female cadets, 90% male. Then, utilizing the item analysis data, a final selection of 110 items for each of two test forms was made, attending to item validity, content, difficulty, and minimization of black-white, male-female, differences in item properties.

The two test forms were administered to MS-II sophomore cadets at all ROTC detachments in two waves, one in the spring of 1982, one in the next MS-II class in the fall of 1982. Data from the two administrations were merged, weighted to consist of 15 to 20% black cadets, 10% women, and to represent the 1980 national distribution of SAT total scores, and analyzed to derive norms as well as certain descriptive statistics about test readability, difficulty, and reliability. In addition, a separate sample was drawn of high school seniors, weighted to conform to the 1980 SAT reference distribution, and analyzed to determine comparability of high school senior performance with that of college sophomores.

The test forms were also administered to samples of graduating senior (MS-IV) ROTC cadets, for each of whom faculty ratings of leadership characteristics and officer potential were obtained. Correlational and regression analyses with the rating criteria were performed for each form separately by gender and ethnic

---

[3]The possible exception to this statement is the Hispanic group. Data came from one very small sample of Hispanic examinees on the island of Puerto Rico, and no adequate estimate could be derived. Additional research may be indicated to check use of the OSB with Hispanic cadets and determine if results of the regression analysis of Form 4 with black cadets was a sampling artifact or characteristic of the test form.

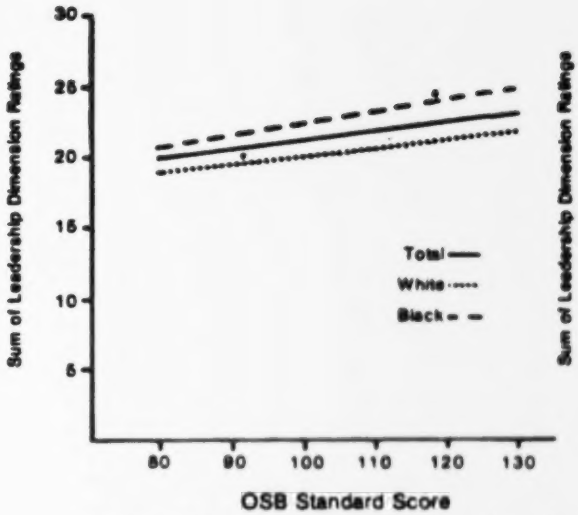


Figure 2. Regression of criterion ratings on OSB scores for black and white samples and for total group--Form 3.

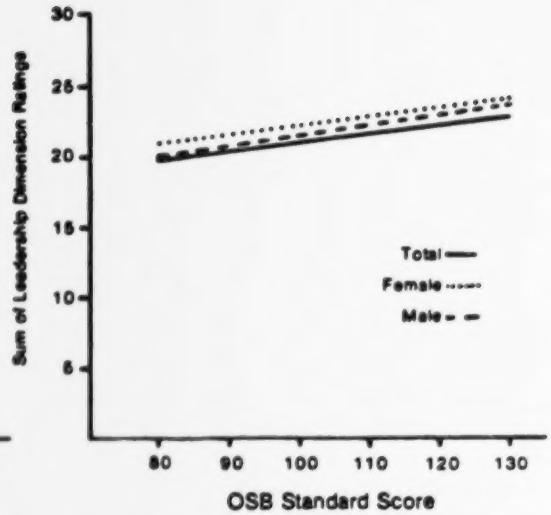


Figure 3. Regression of criterion ratings on OSB scores for male and female samples and for total group--Form 3.

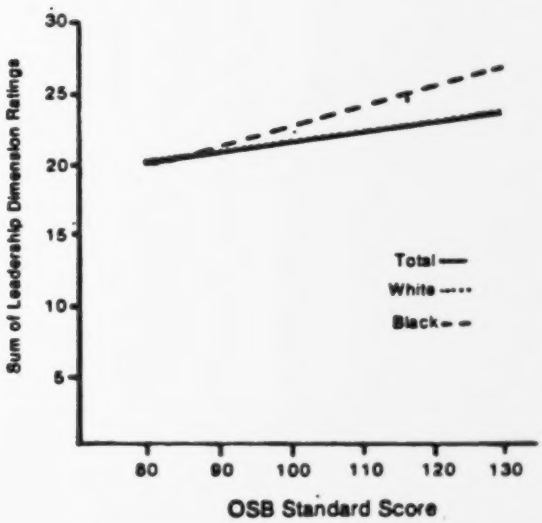


Figure 4. Regression of criterion ratings on OSB scores for black and white samples and for total group--Form 4.

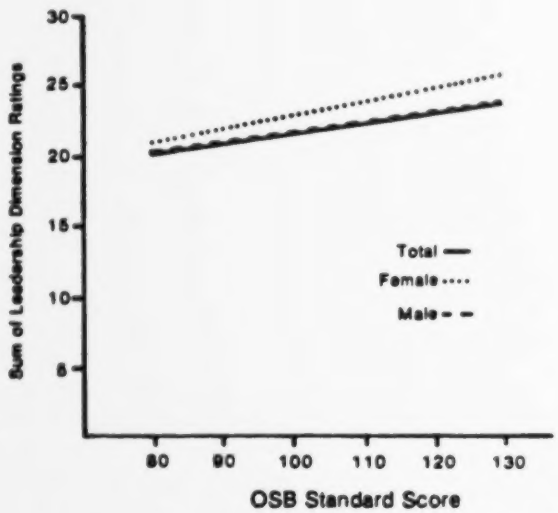


Figure 5. Regression of criterion ratings on OSB for male and female samples for total group--Form 4.

group as well as for the total samples. In addition, one test form was administered to Lieutenants in seven Officer Basic Courses and correlations were obtained with final course grades in each school.

On the basis of the analyses performed the two forms of the Officer Selection Battery may be concluded to be of equivalent job-relevant content and equivalent difficulty, easy to read, with high internal consistency and low standard errors of measurement, and each administrable in two class periods. Separate norm tables were required, and prepared, for each of the two forms. The OSB is more difficult for high school seniors than college sophomores, and tentative norm tables were prepared for the younger group. The OSB is empirically (as well as content) valid, of comparable validity for ethnic and gender subgroups, with no indication of differential validity.

REFERENCES

Clement, S. D., & Ayres, D. B. A matrix of organizational leadership dimensions. Leadership Series Monograph 8. Fort Benjamin Harrison, Ind.: U.S. Army Administration Center, October 1976.

Department of the Army. A review of education and training for officers. Washington, D.C., 1978.

Guion, R. M. Recruiting, selection and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Helme, W. H., Willemin, L. P., & Grafton, F. C. Prediction of officer behavior in a simulated combat situation. ARI Research Report 1182, March 1974. (NTIS No. AD 779 445)

Klemp, G. O., Munger, M. T., & Spencer, L. M., Jr. Analysis of commissioned and non-commissioned naval officers in the Pacific and Atlantic fleets. Boston: McBer and Company, August 1977.

Olmstead, J. A., Cleary, F. K., Lackey, L. L., & Salter, J. A. Development of leadership assessment simulations. (HumRRO Tech. Rep. 73-21) Alexandria, Va.: Human Resources Research Organization, September 1973.

Rogers, R. W., Lilley, L. W., Wellins, R. S., Fischl, M. A., & Burke, W. P. Development of the precommissioning leadership assessment program. ARI Technical Report 560, February 1982.

Sitterson, J. D., Davis, W. P., & Korotkin, A. L. Development of criteria dimensions for evaluation of performance and career development of entry-level officers. Final Contract Report prepared for Army Research Institute. Washington, D.C.: American Institutes for Research, November 1974.

Wellins, R. S., Rumsey, M. G., & Gilbert, A.C.F. Analysis of junior officer training needs. ARI Research Report 1236, February 1980.

Wigdor, A. K., & Garner, W. R. (Eds.). Ability testing: Uses, consequences, and controversies. Washington, D.C.: National Academy Press, 1982.

APPENDIX A

DEFINITIONS OF IDENTIFIED DIMENSIONS OF OFFICER JOB PERFORMANCE

INITIATIVE

Definition: Active attempts to achieve goals, self-starting rather than passive acceptance. Taking actions beyond those called for to achieve goals; originating action. Responding successfully when difficulties arise; identifying and pursuing alternative courses of action if initial approach unsuccessful.

DECISION MAKING

Definition: (a) Problem analysis. Identifying problems, securing relevant information, relating data from different sources and identifying possible causes of problem. (b) Judgment. Developing alternative courses of action and making decisions which are based on logical assumptions and which reflect factual information. (c) Decisiveness. Readiness to make decisions, render judgments, take action or commit oneself.

ADMINISTRATION

Definition: (a) Planning and organizing. Establishing a course of action for self and/or others to accomplish a specific goal; planning proper assignments of personnel and appropriate allocation of resources. (b) Delegation. Utilizing subordinates effectively. Allocating decision making and other responsibilities to the appropriate subordinates. (c) Administrative control. Establishing procedures to monitor and/or regulate processes, tasks or activities of subordinates and job activities and responsibilities. Taking action to monitor the results of delegated assignments or projects.

COMMUNICATION

Definition: Clarity and effectiveness of expression of ideas or desires in writing and orally (formally and informally).

INTERPERSONAL MANNER

Definition: Utilizing appropriate interpersonal styles and methods of influence in guiding individuals or groups toward task accomplishment; accurately appraising the feelings, competence and needs of others; accurately perceiving how viewed by others.

TECHNICAL KNOWLEDGE

Definition: Level of understanding and ability to use technical/professional information.

COMBAT PERFORMANCE

Definition: Effective conduct of combat missions.

APPENDIX B

TEST ITEM TYPES

1. Verbal

   a. Vocabulary
   b. Analogies

2. Quantitative

   a. Quantitative operations
   b. Quantitative reasoning

3. Spatial visualization

   a. Translation of two-dimensional representations into three dimensions
   b. Map reading

4. Problem solving

   a. Problem identification
   b. Information evaluation (judge accuracy, judge relevance, recognize assumptions, evaluate arguments, distinguish between fact and opinion)
   c. Problem analysis
   d. Cause determination
   e. Inductive reasoning
   f. Deductive reasoning

5. General information

   a. Physical, chemical, and biological sciences
   b. Social sciences (history, politics, culture, psychology)
   c. Farm and garden information
   d. "Nature" sports (hiking, fishing, skiing)
   e. Knowledge and application of basic mechanical principles

6. Knowledge about tools, machines, and equipment

7. Interest in technical subjects (mathematics, physical science, mechanical and electronic interests)

8. Interest in rugged, stressing, outdoor subjects (rugged activities, outdoor activities, and military interests)

9. Social problem solving (assertiveness, decisiveness, persuasiveness, competitiveness, task leadership, self-assurance, concern with influencing others, accurate empathy, social leadership, social awareness, social maturity, initiative, persistence, creativity, adaptability, responsibility, resourcefulness)

APPENDIX C

PLAN FOR ADDRESSING EEO FAIRNESS IN OSB

I. Specify Job-Relevant Test Content Domain

a. Analyze junior officer job to obtain job dimensions.
b. Specify content area expected to predict performance on these dimensions.

II. Steps During Item Construction to Avoid Bias

a. Consider each item's relevance to officer's job as well as to test content area.
b. Recognize differences in cultural interpretations of language, vocabulary; avoid penalizing interpretations of particular cultural groups.
c. Avoid items which assume information or understanding that members of certain groups or cultures would not have.
d. Avoid items which demean, stereotype, portray in a negative manner, or otherwise give potential offense to one or more subgroups.
e. Avoid exclusive use of masculine pronoun or implication that all persons in a given category belong to a particular sex, race, or ethnic group.
f. Avoid items which stress values, experiences more familiar to one subgroup/culture than another.
g. Use names, situations, objects which reflect diversity of tested population.
h. Ensure that items will be equally clear to all groups.

   1. Attempt to avoid formal English construction when characteristics of spoken English will provide clearer communication.
   2. Avoid inadvertent use of extraneous information.
   3. Avoid patterns of redundancy which may confuse certain groups.

i. Have items reviewed by knowledgeable, sensitive reviewers to ensure they are consistent with guidelines above.

III. Prepare Test Instructions Which Are Designed to Avoid Bias

a. Take steps noted in II-h above to ensure that instructions will be equally clear to all groups.
b. Read instructions aloud so their comprehension is not dependent solely on reading ability.
c. Give examinees an opportunity to seek clarification of any confusing instructions.
d. Directions should be such that all repetition is symmetric.
e. Avoid stressing negatives in instructions.

IV. Review Instructions for Clarity, Biased Language, and Potential Offensiveness to Particular Groups

V. Item Tryout and Analysis

a. Administer to sample of appropriate proportions of minority and female cadets.
b. Statistical procedure for identification of potential bias.

    1. Determine overall p-value for each item.
    2. Determine p-values for major population groups (males, females, blacks, whites, Spanish-Americans).
    3. Determine differential between overall p-value and each subgroup p-value for each item.
    4. Identify items which have a large p-value differential between overall group and a particular subgroup.

c. Second statistical procedure: Identify items which have particularly low correlations between item performance and criterion score for overall group and for each subgroup.
d. Identify items which have a large item validity differential between overall group and a particular subgroup.
e. For items identified by procedure b, c, or d above, carefully examine for source and revise or reject.

VI. Construct and Standardize Test

a. Select items to satisfy requirements of content, common validity, common difficulty, common response pattern, to the largest extent possible.
b. Administer to sample of appropriate proportions of minority and female cadets.

VII. Validate Test

a. Administer to sample of appropriate proportions of minority and female cadets.
b. Correlate against ROTC criterion.

    1. Total sample.
    2. Separately for major population subgroups.

c. Against on-job performance criterion.
d. Inspect regression lines.

    1. Total sample.
    2. Separately for major population subgroups.

APPENDIX D

WEIGHTING PROCEDURE

A matrix of 21 SAT score intervals by 5 gender/ethnic categories was produced. This 21-row by 5-column table, along with the Revised Target SAT Distribution and the Target Sex/Race Distribution, comprised the data input to the weight calculation program. (See Table D-1 for worksheet.)

The weight calculation program asks the user for the number of rows and the number of columns (21 and 5 in the example) in the matrix and then prompts the user to enter each observed (or combined) cell frequency, one cell at a time. Finally, the program requests the user to input each target row percentage and each target column percentage. The only other value required is the desired total number of weighted cases. If the number of observed cases were entered at this time, the final average weight would be equal to 1.00. In order to use rounded whole-number weights, it is generally best to enter a number several times larger than the observed number of cases. For the example described here, if the number 15000 were entered as the desired total number of weighted cases, the final average weight assigned to each case would be equal to 15000/ 2805 or 5.35. Naturally, some weights will be larger and some smaller than this average.

The most commonly used weighting procedure for a two-dimensional array can be described by the equation:

$$W^{ij} = \frac{R_i \, C_j \, T}{O_{ij}} \qquad (1)$$

where:

W_{ij} = weight to apply to observed cases in cell of row i, column j;

O_{ij} = observed number of cases in cell of row i, column j;

R_i = target proportion for row i;

C_j = target proportion for column j;

T = desired total number of weighted cases.

This procedure has the advantage that all the weights can be calculated in a single step. However, it has the disadvantage that each of the rows is forced to be exactly proportional to all the other rows and each of the columns is forced to be exactly proportional to all of the other columns. While this may be a valid null hypothesis for use in a Chi-square test (the procedure for calculating the expected Chi-square frequency for a cell is directly analogous), it doesn't often match reality.

Table D-1

Weighting Procedure Worksheet: Observed (Combined) Cell Frequencies

		1	2	3	4	5	
Interval number	SAT total interval	White males	Black males	Other males	White females	Black females	Other females
1	400-420						
	430-470	6	0	1	1	4	
2	480-520	8	12	2	1	7	
3	530-570	10	14	0	1	8	
4	580-620	9	19	3	3	6	
5	630-670	45	45	5	9	39	
6	680-720	39	48	7	17	27	
7	730-770	93	39	13	8	14	
8	780-820	154	36	12	17	15	
9	830-870	159	17	9	22	7	
10	880-920	162	11	3	25	4	
11	930-970	220	24	19	47	9	
12	980-1020	215	8	9	32	2	
13	1030-1070	195	6	13	24	3	
14	1080-1120	148	3	5	21	0	
15	1130-1170	116	0	12	24	0	
16	1180-1220	95	3	5	17	1	
17	1230-1270	66	1	4	12	2	
18	1280-1320	33	0	4	12	0	
19	1330-1370	39	0	2	5	1	
20	1380-1420	23	0	1	3	1	
21	1430-1470						
	1480-1520						0
	1530-1570	14	0	0	2	0	0
	1580-1600						0

The alternative used in the research described here consisted of an iterative procedure which alternately weights the rows and then the columns of the matrix. This procedure is continued until the row and column weighted marginal totals stabilize at the target values.

For the first pass through the matrix the following formula is used:

$$E'_{ij} = \frac{R_i \quad T}{\sum_{1}^{i} O_i} O \qquad (2)$$

where:

E'_{ij} = estimated weighted number of cases in cell of row i, column j;

R_i = target proportion for row i;

T = desired total number of weighted cases;

O_{ij} = observed number of cases in cell of row i, column j;

$\sum_{1}^{i}$ = sum of all observed cases in row i.

Application of this procedure to all the rows of an observed matrix will result in a new estimated matrix containing the desired total number of weighted cases (T), and for which the proportion of weighted cases in each row is equal to the target row proportion. The weighted column proportions do not match the target column proportions, however.

Therefore, starting with the E matrix calculated above, apply the following formula:

$$E'_{ij} = \frac{C_j \quad T}{\sum_{1}^{i} E_{ij}} E_{ij} \qquad (3)$$

where:

E'_{ij} = estimated weighted number of cases in cell at row i, column j;

C_j = target proportion in column j;

T = desired total number of weighted cases;

E_{ij} = estimated weighted number of cases in cell at row i, column j calculated in the previous step;

$\sum_{1}^{i} E_{ij}$ = sum of all estimated cases in column j using the E_{ij}s calculated in the previous step.

Now the column proportions will match the target column proportions, but the row proportions will be off. However, the row proportions will be closer to the target row proportions than they are initially. Therefore, apply the following formula:

$$E'_{ij} = \frac{R_i \; T}{\sum_i E_{ij}} E_{ij} \qquad (4)$$

where the definitions of the terms are as indicated above.

Now continue to apply Formulas 3 and 4 alternately until the row proportions and the column proportions both stabilize on the target values. (For the 21 by 5 matrix described earlier, this required 10 applications of each formula.)

The weight to actually be applied to each case ind a given cell is determined by:

$$W_{ij} = \frac{E_{ij}}{O_{ij}} \qquad (5)$$

where:

W_{ij} = weight for cell ij;

E_{ij} = estimated weighted number of cases in cell ij as calculated in the last iteration of the procedure;

Oi_j = observed number of cases in cell ij.

The program used automatically carries out 15 iterations of the procedure and then prints out the weights.

NOTE: In order to use integer (whole number) weights, it is recommended that, if many of the calculated weights are so small that rounding them may have a large effect, the weighting program should be rerun using a larger value for the desired total number of weighted cases. The new set of weights will be exactly proportional to the old set, but rounding will have a smaller effect. Cells with no cases in them will naturally have a weight of zero.

The note above was observed for the 21 by 5 matrix used in this example, and a desired total weighted sample size of 40,000 was entered. Since there were 2805 observed cases, the average weight was approximately 14 and the range was from a high of 98 (for cell 1,1, white males with very low SAT total scores) to a low of 2 (for cells 19,5 and 20,5, Other females with high SAT total scores).

APPENDIX E

READABILITY MEASURES

A. Flesch Reading Ease. This is probably the most widely known formula. Its popularity is due to its ease of computation, its correlation with more complex (and presumably more accurate) measures and, of course, because of the numerous publications of its author. The formula and the interpretation of the results follow.

Reading Ease = 206.835 - 0.846 w1 - 1.015 s1

w1 = number syllables per 100 words
s1 = average number words per sentence

90-100 = Very Easy
80-90 = Easy
70-80 = Fairly Easy
60-70 = Plain English
50-60 = Fairly Difficult
30-50 = Difficult
0-30 = Very Difficult

B. Sticht. This measure is simply the percentage of 1-syllable words used in a passage. It correlates highly with reading comprehension as measured by a Cloze Test (readability of a passage with every nth word deleted). Sticht developed this measure for the Army, and then refined it, whereupon it became the FORCAST measure.

C. FORCAST. This measure was developed as part of an extensive study of readability of Army materials. It is an acronym made up of the names of the developers: FORd, CAylor, and STicht.

$$\text{Grade Level} = 20 - \frac{\text{Number of 1-syllable words in 150 words}}{10}$$

This formula results in a grade-level equivalent, and was developed on samples of Army personnel and samples of Army prose. It thus may be highly relevant for the present investigation.

D. The Gunning Fog Index. This index estimates the reading grade level required for understanding the material. It is based upon text from various American magazines fromn "pulp" to "class" and from passages in the McCall-Crabbs reading measures for grades 6 to 12.

Grade Level = 0.4 (Average words per sentence + number of words with $\geq$ 3 syllables)

E. Kincaid/Flesch. This formula was derived from a study of Navy enlisted personnel. The reading grade level estimated by the formula is the level at which 50 percent of the subjects could correctly fill in 40 percent of the blanks on a Cloze Test.

Grade Level = 0.39 (Average sentence length) + 11.80 x
(average number syllables per word) - 15.59

F. Flesch (Original). In 1949 Flesch produced a chart on which scales showing words per sentence and syllables per 100 words appear in vertical columns. By using a ruler to align the values of these two variables in a given reading passage, the user can obtain the reading ease score from the point at which the ruler intersects a third vertical scale centered between the other two.

APPENDIX F

COMPARISON OF OFFICER SELECTION BATTERY WITH TWO TESTS IN USE AS SELECTORS FOR ARMY OFFICER PRECOMMISSIONING TRAINING PROGRAMS

Background

Forms 3 and 4 of the Officer Selection Battery (OSB) were developed to replace subtest 2 of the Cadet Evaluation Battery (CEB), used to select applicants for Advanced ROTC or entrance to Officer Candidate School (OCS). In OCS the CEB is designated OSB Forms 1 and 2.

The SAT is required by many colleges and universities as a selection measure. Thus SAT scores were available on some ROTC cadets. This Appendix describes the results of certain statistical comparisons of the OSB, SAT, and CEB.

Descriptive Statistics

Because the number of women and ethnic minorities with complete data (i.e., an OSB and CEB score or SAT score, and a criterion rating) was relatively small, the OSB-3 and 4 samples were combined for most of the analyses.

A. The Sample--A Caveat

The sample of MS-IV cadets with SAT scores is shown in Table F-1. This sample has a number of incidental qualities to it. First, it probably does not include any cadets from schools with open enrollment policies, the result of which would have been an underrepresentation of lower ability students. Second, the SAT tends to be required for admission by the more competitive schools. Thus the scores of the sample of cadets used in this analysis would be expected to be superior to the MS-IV total sample, some of whom qualified on CEB or certain other tests. Third, there were only 36 black cadets in the sample, and 12 of the 36 were students at a single, historically black, college. Finally, there were only 10 Hispanic cadets in the sample with SAT scores, so those cadets were not included in many of the analyses reported here.

The data presented in the analyses that follow should be interpreted with respect to the incidental quality of the sample, and should not be considered generalizable to a population of college students or ROTC cadets.

B. The Criterion

Table F-2 shows the means and standard deviations on the two criterion ratings used in the research: the rating on Officer Potential and the sum of the ratings on six dimensions of leadership performance.

Table F-1

Means and standard deviations of MS-IV cadets with OSB and SAT

	SAT/OSB Sample		
	N	Mean	Standard Deviation
Total sample:			
OSB	301	114.8	20.0
SAT	301	995.2	192.9
Males:			
OSB	267	114.5	20.1
SAT	267	989.5	192.3
Females:			
OSB	34	116.7	19.7
SAT	34	1,042.1	193.9
Blacks:			
OSB	36	93.2	16.1
SAT	36	750.8	173.8
Whites:			
OSB	246	118.0	18.6
SAT	246	1,029.8	170.4

Table F-2

Means and standard deviations on cadet rating form for MS-IV cadets who took both OSB and SAT

	N	Mean	Standard Deviation
Total:			
Officer Potential	301	3.7	1.0
Sum of Leadership Dimensions	301	22.7	5.4
Males:			
Officer Potential	267	3.7	1.0
Sum of Leadership Dimensions	267	22.6	5.3
Females:			
Officer Potential	34	3.9	1.1
Sum of Leadership Dimensions	34	23.9	6.1
Blacks:			
Officer Potential	36	3.3	1.2
Sum of Leadership Dimensions	36	19.9	5.3
Whites:			
Officer Potential	246	3.7	1.0
Sum of Leadership Dimensions	246	23.0	5.3

The difference between the means of blacks and whites for the rating on Leadership Dimensions is statistically significant ($p < .01$), with the ratings for blacks being about three points lower than for whites. The race of the raters was not requested of them, so any moderator effect of that variable cannot be tested. The majority of the 36 black cadets were students at historically black colleges, so most blacks were thus usually compared with other blacks, whether rated by black or white raters.

C. The Predictors

The means and standard deviations of the two predictors, based upon data from identical samples, are shown in Table F-3.

Table F-3

Means and standard deviations of MS-IV cadets with OSB and SAT scores

	SAT/OSB Sample		
	N	Mean	Standard Deviation
Total Sample:			
OSB	301	114.8	20.0
SAT	301	995.2	192.9
Males:			
OSB	267	114.5	20.1
SAT	267	989.3	192.3
Females:			
OSB	34	116.7	19.7
SAT	34	1,042.1	193.9
Blacks:			
OSB	36	93.2	16.1
SAT	36	750.8	173.8
Whites:			
OSB	246	118.0	18.6
SAT	246	1,029.8	170.4

The test performance of black cadets is significantly lower than that of white cadets on both predictors, and the difference is about the same as in the total MS-IV sample. There is an interesting reversal in this pattern, however, in the male/female comparisons. In the total MS-IV sample, the OSB mean for males was almost eight points higher than that for females. In the OSB/SAT sample the OSB mean for males is about two points lower than that for females. The screening on SAT appears to have affected the female subsample more than other subsamples.

Validity

Table F-4 presents the validities of the OSB and SAT, for the total sample and by subsample, separately for Forms 3 and 4.

Table F-4

Validity of OSB and SAT, by subgroup, for MS-IV sample with both OSB and SAT scores
(Validity of OSB total sample of MS-IV in parentheses)

	N	OSB		SAT	
		Officer Potential	Leadership Dimension Sum	Officer Potential	Leadership Dimension Sum
Form 3					
Black	17	-0.07(0.21)	0.02(0.27)	-0.09	-0.23
White	103	0.04(0.20)	0.10(0.24)	0.04	0.11
Hispanic	3	-0.60(0.10)	-0.44(0.33)	-0.22	-0.03
Female	9	-0.17(0.28)	-0.19(0.30)	0.50	0.44
Male	119	0.06(0.19)	0.16(0.26)	0.03	0.12
Total	128	0.04(0.20)	0.14(0.26)	0.05	0.14
Form 4					
Black	19	0.71(0.39)	0.71(0.34)	0.76	0.73
White	143	0.21(0.26)	0.23(0.23)	0.24	0.28
Hispanic	7	0.14(0.26)	0.19(0.31)	0.49	0.53
Female	25	0.43(0.32)	0.46(0.33)	0.57	0.56
Male	148	0.27(0.27)	0.29(0.25)	0.32	0.35
Total	173	0.30(0.28)	0.32(0.28)	0.37	0.40

Results for the total, white, and male samples are the only ones in Table F-4 with enough cases to warrant any interpretation, and of course, many of the cases in these subsamples overlap.

Validity of both tests is higher in the portion of the sample administered Form 4. Since the SAT validity is totally independent of the OSB, it must be some underlying sampling difference, not test difference, that affects the validity of these two separate cognitive tests.

In order to increase the number of female and black cadets for analysis, the Form 3 and Form 4 samples were combined, and intercorrelations and validity coefficients obtained for the total sample, and for females and blacks separately. The correlations for whites and males were not recomputed since they would essentially duplicate those of the total sample. Table F-5 presents the matrices.

Table F-5

Intercorrelations among OSB, SAT, and criterion measures for MS-IV cadets who took both tests

	OSB 3/4	SAT	Officer Potential	Leadership Dimension Total
Total Sample (N=301)				
OSB 3/4	1.00	0.78	0.17	0.23
SAT		1.00	0.21	0.28
Sum of Leadership Dimensions			1.00	0.92
Females (N=34)				
OSB 3/4	1.00	0.82	0.22	0.25
SAT		1.00	0.48	0.48
Sum of Leadership Dimensions			1.00	0.97
Blacks (N=36)				
OSB 3/4	1.00	0.85	0.36	0.41
SAT		1.00	0.44	0.39
Sum of Leadership Dimensions			1.00	0.91

In every comparison except one in these samples the SAT was more valid against both criteria than was the OSB. The difference was small for the total sample, four or five correlation points, but substantial for the sample of 34 women. Among the 36 black cadets a small difference in validity favoring SAT is seen against one criterion, and favoring OSB against the other.

Table F-6 presents validity coefficients for the OSB and SAT at all individual schools with more than one SAT score. In the group of nine schools with ten or more cases, the OSB is more valid in four schools and the SAT is superior in five. Obviously this analysis is inconclusive because of the small size and questionable characteristics of the samples.

Table F-6

School by school validity coefficients for OSB and SAT
(Analysis limited to individuals who have both OSB and SAT)

		OSB		SAT	
School #	N	Officer Potential	Leadership Dimension Total	Officer Potential	Leadership Dimension Total
1	3	0.91	0.75	1.00	0.92
2	5	-0.34	-0.37	-0.05	-0.09
3	7	0.23	0.45	-0.13	0.10
4	9	0.69	0.76	0.73	0.81
5	3	-0.06	-0.43	-0.30	-0.64
6	15	0.08	0.10	-0.03	0.05
7	3	0.43	0.26	-0.74	-0.60
8	12	0.66	0.70	0.70	0.75
9	5	-0.25	0.01	-0.11	-0.09
10	3	-1.00	-1.00	0.76	0.76
11	5	-0.98	-0.87	-0.89	-0.91
12	2	1.00	1.00	1.00	1.00
13	10	0.22	0.31	0.17	0.19
14	25	0.09	0.30	0.14	0.32
15	2	0.00	-1.00	0.00	-1.00
16	2	0.00	0.00	0.00	0.00
17	16	0.18	0.26	0.36	0.61
18	17	0.54	0.60	0.49	0.56
19	6	0.39	0.38	0.46	0.39
20	8	-0.07	0.07	0.06	0.19
21	23	0.04	0.04	0.10	0.10
22	3	0.68	0.61	0.81	0.76
23	4	-0.83	-0.71	-0.61	-0.59
24	21	0.15	0.38	0.19	0.42
25	21	0.14	0.06	0.12	0.04
26	5	0.13	-0.13	-0.75	-0.74
27	2	-1.00	-1.00	-1.00	-1.00

Cutting Scores

Two sets of decision tables were prepared, using data from the MS-II sample tested for the standardization of the OSB. Table F-7 presents the consequences of various qualifying scores among cadets who had taken both OSB and CEB. Table F-8 presents the same analysis, among cadets with both OSB and SAT scores.

In the first group, all MS-II students with OSB and CEB scores, the OSB would screen out larger percentages of female and black cadets at every score level than would the CEB. This, of course, is also the case with cadets in general. Fewer qualify on OSB than on CEB.

Alternatively, if the SAT were used to select Advanced Course cadets, from the sample with both OSB and SAT scores, results in both the total group and subgroups would be much more similar. In the total group the percentages screened out by OSB or SAT would be virtually identical, with not much difference in the percentage of black cadets screened out by either test, and only a slight, but consistent, difference among female cadets.

Prediction of Officer Basic Course Grades

When OSB-3 was administered to 577 officers attending seven Officer Basic Courses (OBC), the results were high validity coefficients between test scores and final course grades. This led to the question of how well the SAT and CEB would correlate with OBC final course grades. SAT scores were available for 61 of the officers in that sample, and CEB scores were available for 222 of the officers. The correlations are presented in Table F-9.

Table F-7

Percentage of individuals in various subgroups eliminated by various minimum scores
(Analysis limited to MS-II cadets who have both OSB and CEB scores)

OSB Score	Total	Males	Females	Blacks	Whites	HBC	CEB Score	Total	Males	Females	Blacks	Whites	HBC
80	16	13	29	46	4	50	86	15	14	23	37	8	37
85	23	20	38	63	8	66	90	21	19	32	50	11	50
90	31	27	45	75	14	78	94	27	25	39	60	15	59
92	34	30	49	78	16	81	96	33	30	46	69	20	69
95	40	36	56	84	22	88	98	39	36	54	76	26	76
97	43	39	59	87	25	90	100	40	36	54	76	26	76
100	48	44	64	90	31	93	102	46	42	61	82	32	83
N	3911	3150	758	913	2703	769		3911	3150	758	913	3703	769

Table F-8

Percentage of individuals in various subgroups eliminated by various minimum scores
(Analysis limited to MS-II cadets who have both OSB and SAT scores)

OSB Score	Total	Males	Females	Blacks	Whites	HBC	SAT Score	Total	Males	Females	Blacks	Whites	HBC
80	5	4	11	32	1	41	620	4	3	10	30	1	42
85	8	6	17	45	3	57	690	7	6	14	45	2	60
90	12	11	22	57	6	68	760	12	10	20	59	5	75
92	15	13	26	62	8	76	790	14	12	21	63	79	77
95	19	18	29	69	13	83	830	18	17	25	71	11	82
97	22	20	32	74	15	85	850	21	20	28	73	14	82
100	27	25	38	80	20	89	880	25	23	32	79	17	86
N	2914	2453	456	312	2459	182		2914	2453	456	312	2459	182

Table F-9

Correlations with final grade in Officer Basic Courses

Sample		N	Correlation Coefficient
All officers with OSB and CEB; all courses combined:			
	OSB	222	0.48
	CEB	222	0.36
All officers with OSB and SAT; all courses combined:			
	OSB	61	0.50
	SAT	61	0.57
All officers with OSB, CEB, and SAT; all courses combined:			
	OSB	55	0.49
	CEB	55	0.45
	SAT	55	0.56

870212